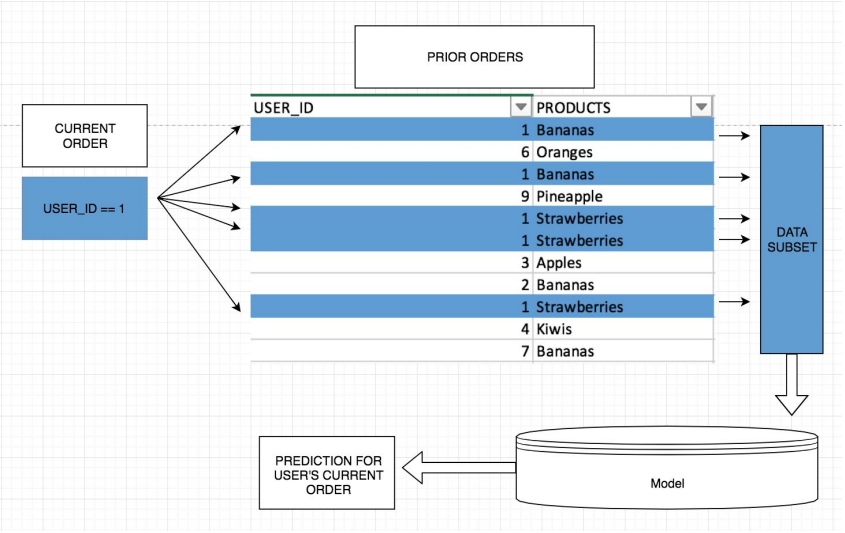
Group E: Anthony Rentsch, Pei Gong, Jenn Halbleib
STAT 495: Advanced Data Analysis
December 18th, 2017

Final Project Executive Summary

Description of the Competition: Instacart is an online personal shopping service focused on delivering users products from local retailers. Instacart started this Kaggle competition with the goal of predicting products users will reorder. The data set consists of lists of products in each order, what order number this is for the user, time of day, and day of week.

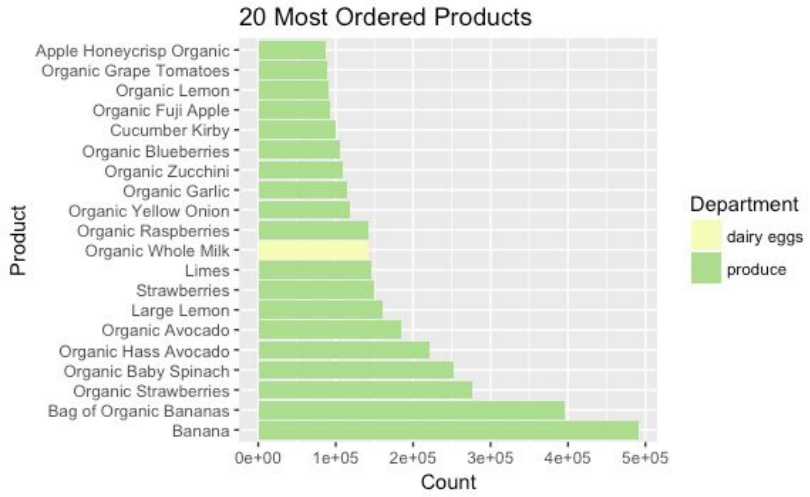
Ultimate Approach: With little experience working with a strictly historical data set like this one, we embraced the concept of an interpretable minimum viable product. For each user, we simply looked at their order history and predicted that they would order the product(s) that they had ordered most often. We predicted that a user would order the *1, 2, 3, and 4 product(s)* they have ordered most often, with ties broken at random. We observed slight improvement as we considered more products (F1 scores: 0.163, 0.219, 0.243, 0.214, respectively), although we would not expect to keep improving our score as we considered more items, as the metric we were evaluated on (F1 score) penalizes wrong guesses. For comparison, the top Kaggle score was 0.409.



Some "Failed" Attempts:

Bananas: In our first pass at the data, we investigated the raw frequency of products ordered per hour, per day of the week, and in general. We were surprised to discover this inquiry did not give us much usable information, since the most ordered product in every case was bananas. To get a

baseline model, we predicted every user in the test ordered bananas and obtained a Kaggle score of 0.035.



CART: After the improvement in scores we obtained in our ultimate approach described above, we started to think about ways to use methods we had used in class to hopefully improve our predictions. We decided to try fitting a CART tree to each user's prior orders to predict the items in their current order. This approach worked poorly, due a lack of user specific information on which to build the tree. With this method, we earned an F1 score of 0.

Future Ideas:

Users: To improve on our predictions, we thought about generalizing each user's preference into categorical variables. For example, one third of the users have ordered organic products in the past. So we can code a variable "organic" specific to each user, which takes a value of 1 when users have ordered organic products in the past and a value of 0 otherwise. In similar fashion, we can code variables such as "vegetarian," "family with baby," and "meat lovers."

Network: We tried expressing our data as a network: products ordered in the same order are "connected" to one another. Then, the number of connections between any two products could be used to to predict the probability that both products end up being ordered if at least one is ordered. However, the data we have does not explicitly say what products are in the cart already, so we would have to predict what items begin in the order and then use a network to predict subsequent items. We could also consider an approach in which users are "connected" to other users if they order the same products and use collaborative filtering, although that is outside the scope of our current project.

Make a Decision Tree: Obviously not every product ordered is a reordered product. So, making a decision tree that predicts "reorder" or "new order" at the top is one way to explore making better predictions with this data set. This could extend to a Bayesian Hierarchical Model.